DOCUMENT RESUME

ED 298 148 TM 012 238

AUTHOR Carlson, Robert E.
TITLE Alternative Procedures for Collecting and Analyzing Job-Relatedness Judgments in NTE Test-Appraisal Studies.
PUB DATE Apr 88
NOTE 19p.; Paper presented at the Annual Meetings of the American Educational Research Association (New Orleans, LA, April 5-9, 1988) and the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Content Validity; Item Analysis; Job Performance; *Licensing Examinations (Professions); *National Programs; Research Methodology; Scores; *Teacher Certification; Teacher Education; Test Theory; *Test Validity
IDENTIFIERS *Job Relatedness; *NTE Core Battery; Test Appropriateness

ABSTRACT

Early National Teacher Examinations test-appraisal studies focused only on the extent to which the content covered on a test matched the content of teacher education curricula. Recently, there has been a shift in emphasis from adequacy-of-preparation, which is now seen by many as irrelevant to establishing the appropriateness of a teacher licensure examination, to job-relatedness. The procedures used to collect and analyze job-relatedness judgments in many studies, however, seem designed to yield positive results. Alternative procedures are described: (1) focusing the judgments to be made on necessity for successful job performance, rather than on relevance to the job; (2) asking for a total-test judgment regarding how well the tested content represents the entire domain of necessary content; and (3) computing an average per-item score, rather than a majority index. These procedures pose a more rigorous test of a licensure examination's job-relatedness. They incorporate the legal and professional requirements of appraising licensure tests, and they yield richer, more meaningful data regarding the appropriateness of using a test as a licensure requirement. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298148

ALTERNATIVE PROCEDURES FOR COLLECTING AND ANALYZING JOB-RELATEDNESS JUDGMENTS IN NTE TEST-APPRAISAL STUDIES*

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Robert E. Carlson
IOX Assessment Associates

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ROBERT E. CARLSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Background and Purpose

The NTE Program tests, a battery of tests formerly called the National Teacher Examinations, are the most frequently used tests for teacher licensure in this country. Produced by the Educational Testing Service (ETS), NTE tests are or soon will be used in 22 states as a requirement for teacher certification (U.S. Department of Education, 1987). Because ETS requires that state agencies using the tests for teacher licensure conduct studies to "establish that the tests are a valid measure of the knowledge and skills required for the intended use" (NTE Policy Council, 1985, p. 9), NTE tests have received substantial scrutiny in recent years. At least 35 studies of NTE tests have been conducted, most of them by ETS (Cross, 1985).

The primary activity in NTE test appraisals has been the collection of content-related evidence of validity. As described in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985), "content-related evidence demonstrates the

*A symposium presentation at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 5-9, 1988.

TM012 238

degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content" (p. 10).

In the early NTE test-appraisal studies conducted by ETS, the "universe or domain of content" was defined exclusively in terms of the curricula of teacher education institutions (e.g., ETS, 1976a, 1976b, 1979). Judgments were sought from faculty at teacher training institutions regarding the match between the tests' content and the institutions' curricula. Perhaps because such judgments were used in court to defend successfully South Carolina's use of NTE tests for teacher licensure (United States v. South Carolina, 1978), the collection of what has come to be known as adequacy-of-preparation data became a central feature of most subsequent NTE test-appraisal studies.

The publication in 1978, however, of the Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice, 1978) led to a reconceptualization of the "universe or domain of content" appropriate for appraisals of NTE tests. According to these federal guidelines, the universe of content should consist of the knowledge and skills that are necessary for successful performance on the job for which the candidates are being evaluated.

The first studies to investigate the job-relatedness of NTE tests were conducted in 1982 in Arkansas (Roth, 1982) and Virginia (Cross, 1982). Public school teachers and teacher

educators were asked to judge the relevance of the tests' content to the job of teaching. Since then, most studies have looked at both adequacy-of-preparation and job-relatedness (e.g., Bowman, 1984; Faggen, 1983; IOX Assessment Associates [IOX], 1983; Poggio, 1985; Radbil, 1983). This "dual" approach has been endorsed by the NTE Program Policy Council in their Guidelines for Proper Use of NTE Tests (1985).

Recently, however, it has been argued that job-relatedness should be the primary focus of NTE test-appraisal studies (e.g., Cross, 1988; Mehrens, 1987; Rebell, 1986). Because teacher licensure tests are designed to protect the public from incompetent teachers, judgments about the tests' appropriateness should be based on the extent to which the tests cover the knowledge and skills necessary for successful job performance. According to this view, adequacy-of-preparation judgments are irrelevant.

A test-appraisal study that reflects this view was recently completed for the state of South Carolina (IOX Assessment Associates, 1988). In that study of three NTE tests, judgments regarding the tests' job-relatedness were collected. No judgments were solicited regarding adequacy-of-preparation.

Thus, it may be that in future NTE test-appraisal studies, the primary type of content-related evidence of validity collected will be judgments about the job-relatedness of the tests' content. The purpose of this paper is to describe a methodology for collecting and analyzing job-relatedness

judgments that is more rigorous than that typically employed in NTE test-appraisal studies.

The Commonly Used Job-Relatedness Question

Among the earliest NTE test-appraisal studies to incorporate job-relatedness judgments were those conducted by Cross (1982) and McMillan (1983) in 1982 for the Commonwealth of Virginia. In these studies, teacher educators as well as public school teachers and instructional supervisors were asked to judge "the relevance of the content measured by each item with respect to the domain of knowledge an entry-level teacher should possess" (Cross, 1984, p. 3). Response choices were "crucial," "important," "questionable," and "not relevant." The same procedure, with occasional minor changes in the question asked but not in the response choices, has been used in almost all NTE test-appraisal studies since first introduced in 1982 (e.g., Kalisch, 1986; Masonis, 1984; Poggio, 1985; Radbill, 1983).

The language utilized in this procedure, both in the description of the judgment to be made and in the response options, must be viewed in light of the nature of job-relatedness called for in the Uniform Guidelines and the AERA-APA-NCME Standards. According to the Uniform Guidelines, "selection procedures which purport to measure knowledge, skills, or abilities may in certain circumstances be justified by content validity...if that knowledge, skill, or ability is a necessary [italics added] prerequisite to successful job performance" (Section 14C(1)). Similarly, the Standards indicate that a

"rationale should be provided to support a claim that the knowledge or skills being assessed are required [italics added] for competent performance in an occupation" (Standard 11.1, p. 64).

Thus, the knowledge or skills measured by NTE tests should be required for successful job performance, not simply relevant to the job. Clearly, the domain of knowledge and skills relevant to teaching is much broader than the domain that is necessary for competent performance. Therefore, using relevance as a criterion is not a very strict measure of a test's appropriateness.

Furthermore, of the four response choices, only "crucial" can be reasonably equated with "necessary" and "required." The response "important" does not convey the same meaning as "required." Required skills and knowledge are probably only a subset of important skills and knowledge. Yet in many of the studies using this approach, both "crucial" and "important" responses are categorized as positive, that is, as indicating the job-relatedness of an item, further reducing the rigor of the appraisal methodology.

This common question used for collecting job-relatedness judgments may be, at least in part, what led Madaus and Pullin (1987) to assert that "the procedures currently used to establish content validity are biased in favor of positive, rather than disconfirming, answers to questions about the knowledge, skills, and competencies needed to teach at some minimal level of effectiveness" (p. 36).

An Alternative Job-Relatedness Question

In recent NTE test-appraisal studies conducted by IOX Assessment Associates (IOX), a more stringent question has been asked about items' job-relatedness (IOX, 1987, 1988). In a 1987 study of three NTE specialty area tests (IOX, 1988), teachers and teacher educators were asked to respond "yes" or "no" to the following question for each test item:

> Is the knowledge or skill needed to answer this item necessary for satisfactory performance by a beginning South Carolina teacher in the specialty area tested?

The language of this question embodies the criterion of necessity for successful performance as formulated in the Uniform Guidelines and the Standards rather than the more lenient criterion of relevance to the job. It therefore seems to be a more appropriate question to ask regarding the job-relatedness of test items.

The Importance of a Total-Test Question

In addition to item-by-item judgments of job-relatedness, some studies have included total-test judgments (Echternacht, 1985, 1986; IOX, 1983). In these studies, participants were typically asked, after making their individual item judgments, to judge the importance of the totality of the skills and knowledge covered on the test. This judgment, however, essentially provides no new information about the test. It is simply an aggregate judgment of the same nature as the item-by-item judgments.

A different total-test judgment was used by IOX in recent

NTE specialty area test-appraisal studies for South Carolina (IOX, 1987, 1988). In those studies, after making their item-by-item judgments, participants were asked to provide an estimate of the extent to which a test's content was representative of the totality of knowledge and skills needed by a beginning teacher. Specifically, participants were asked the following question:

> What percentage of the subject-related knowledge and skills needed by a beginning South Carolina teacher in the specialty area tested is covered on this test form?
>
> Estimated percentage = _____ %
> (to nearest five percent)

This type of total-test judgment is necessary to assess the degree to which a test's items truly represent the content domain of interest, in this case, the subject-matter related requirements of teaching in the subject field covered by the particular NTE test. Although the vast majority of an NTE test's items may be judged job-related, this does not indicate that those items--in the aggregate--suitably represent the entire domain of knowledge and skills required for successful job performance as a teacher. The test's items, while important in their own right, may fail to tap many important dimensions of that domain.

To illustrate, consider the two schematic representations in Figure 1. In both representations, all the items, individually, have been judged to be job-related. In the illustration on the left, however, the set of items on the test covers only a small portion of the needed knowledge and skills. In the illustration on the right, the job-related items cover far more effectively

the full range of content required by the job.

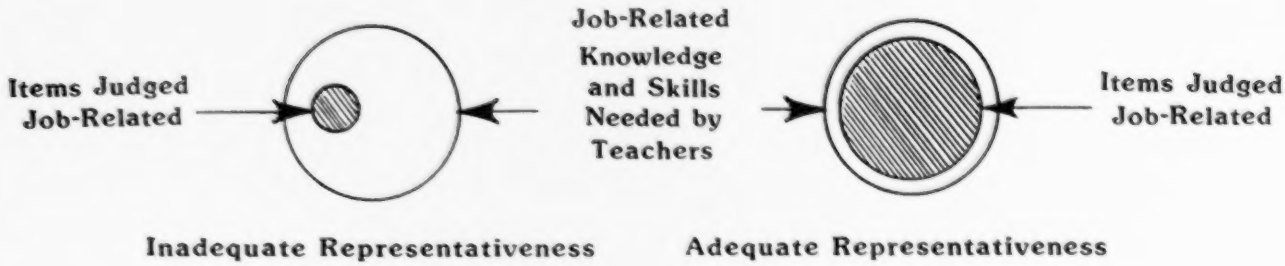

Figure 1. Two contrasting job-representativeness situations, in each of which all test items are judged job-related.

As illustrated in Figure 1, item-by-item judgments of job-relatedness alone are not sufficient to assess how well a test represents the content domain of interest. The importance of this total-test judgment was realized in a recent South Carolina study in which the NTE specialty area test in Spanish was appraised (IOX, 1988). The item-by-item job-relatedness judgments were within an acceptable range, but the mean response for the total-test judgment was only 65%, much lower than the results for most of the other tests for which the question was asked. From participants' comments, it was learned that the low rating was given because the test did not cover the important area of oral proficiency. Had the total-test judgment not been sought, the Spanish test would have appeared to be as appropriate as the other tests.

Data Analysis

A second area in which a more appropriate degree of rigor can be applied in appraising NTE tests is that of data analysis.

The Majority Index

Investigators using the four-point scale discussed above (i.e., crucial, important, questionable, and not relevant) have typically analyzed the results as follows. Although participants select one of four response options, in the analysis the responses are collapsed into two categories: "crucial" and "important" are categorized as "job-relevant," and the remaining two are categorized as "not job-relevant." Each test item is then classified as either job-relevant or not job-relevant based on a majority-rule standard. That is, a test item is classified as job-relevant if more than 50% of the participants rated the item job-relevant (i.e., "crucial" or "important"). Finally, the percentage of items on a test that have been classified as job-relevant is computed. It is this "majority index" that is reported.

This analysis procedure clearly seems biased in favor of positive results. First, participants use four response options, but the analysis uses only two. It's like asking one question and then using the results as an answer to a different question. If responses are treated as "job-relevant" or "not job-relevant," why not allow participants to select from those two options? Second, the key discrimination point between an item being considered job-relevant or not is the distinction between

"important" and "questionable," a choice that probably inclines participants to opt for the more positive alternative. Finally, the majority index does not allow any discrimination among items in terms of their judged job-relevance beyond the yes/no distinction. Items that are judged positively by 51% percent of the reviewers are regarded as equally acceptable as items judged positively by all reviewers.

Some investigators tip the scale even more in favor of positive responses by conducting a second analysis in which judgments of "questionable" are distributed among the other response options. This was done in a study of selected NTE tests for the state of New York (Faggen, 1985). The "questionable" responses for each item were distributed among the two major classifications, that is, job-relevant and not job-relevant, proportionately to the way the other responses were distributed for that item. For example, assume that an item receives the following distribution of responses:

Crucial	-	4
Important	-	5
Questionable	-	8
Not Relevant	-	3

The total number of "crucial," "important," and "not relevant" responses is 12. Of those, nine, or 75%, are in the job-relevant category (i.e., "crucial" or "important"). Therefore, 75% of the "questionable" responses, or six, are added to the job-relevant category, and the remaining two "questionable" responses are added to the not job-relevant category. The justification offered for this procedure is that "some uncertainty exists as to

why panel members chose [the 'questionable' category]" (Faggen, 1985, p. III.26). The effect of this procedure can be significant: On one test, the percentage of items judged job-relevant increased from 82.5% to 98.3% when this procedure was employed (Faggen, 1985, Tables III.14 and III.19A).

The Average Per-Item Score

An alternative way to analyze job-relatedness data has been used in studies conducted by IOX Assessment Associates (IOX, 1983, 1987, 1988). This procedure summarizes the yes/no data obtained in response to the job-relatedness question presented earlier on page 6. For each test item, the percentage of "yes" responses is computed, yielding a "per-item score." Next, an "average per-item score" for the total test is computed by adding the per-item scores then dividing the sum by the number of items on the test. This procedure is preferable to the majority index because (a) no post-data-collection collapsing of responses is involved, and (b) a test item's job-relatedness can be quantified on a scale from 0 to 100, rather than on a dichotomous job-relevant/not job-relevant scale.

As would be expected given the different ways in which the majority index and the average per-item score are determined, they can yield quite different results, and the difference tends to increase as the average per-item score decreases. This is illustrated in Table 1, which shows the results for three NTE specialty area tests appraised in South Carolina. The job-

relatedness question on page 6 was used, and both the average per-item score and the majority index were computed.[1]

Table 1

Comparison of Majority Index and Average Per-Item Score for Selected NTE Specialty Area Tests

Test	Average Per-Item Score	Majority Index
Biology/General Science[a]		
Form A	93.1	100.0
Form B	93.0	99.4
French[a]		
Form A	80.4	92.8
Form B	75.7	87.2
Agriculture[b]		
Form A	78.6	94.7
Form B	78.4	95.3

[a] Data from IOX, 1987. [b] Data from IOX, 1988.

In each case, the majority index gives a more positive result than the more data-sensitive average per-item score. The difference is most dramatic for the agriculture test. The average per-item scores for the two test forms reviewed are about 78, suggesting some lack of job-relatedness. According to the majority indices, however, which are about 95, the test appears to be very job-related.

Summary and Conclusion

Early NTE test-appraisal studies focused only on the extent to which the content covered on a test matched the content of teacher education curricula. The publication of the Uniform Guidelines in 1978 led to the collection of job-relatedness judgments in such studies in addition to judgments about the match between the test and the curricula. In the last few years, there has been a shift in emphasis from adequacy-of-preparation, which is now seen by many as irrelevant to establishing the appropriateness of a teacher licensure examination, to job-relatedness (Coyle, 1988). Recently, in some NTE test-appraisal studies, only job-relatedness judgments have been collected. As more attention is focused on the issue of job-relatedness, it is important that the procedures used to collect and analyze job-relatedness data be appropriate for the task. That is, the procedures should be fair but rigorous, and should allow for maximum confidence in the decision to be made regarding whether or not a test is suitable for use to license teachers.

The procedures used to collect and analyze job-relatedness judgments in many past studies seem designed to yield positive results. At each of a series of minor choice-points in the procedures, the choice typically made has the effect of enhancing the judged job-relatedness of a test.

Alternative procedures have been described in this paper. They involve the following:

1. focusing the judgments to be made on necessity for successful job performance, rather than on relevance to the job,

2. asking for a total-test judgment regarding how well the tested content represents the entire domain of necessary content, and

3. computing an average per-item score, rather than a majority index.

These procedures clearly pose a more rigorous test of a licensure examination's job-relatedness. Given the seriousness of the decisions to be made based on examinee scores, that is, whether or not a person will be allowed to pursue a career in public education, such strict procedures seem warranted. Although the proposed procedures are more demanding, they do not seem to be unfair. Rather, they incorporate the legal and professional requirements for appraising licensure tests, and they yield richer, more meaningful data regarding the appropriateness of using a test as a licensure requirement.

Footnotes

1. In NTE test-appraisals, it is never clear how good is good enough. A test's judged job-relatedness should be reviewed in relation to the judged job-relatedness of other tests. To allow for comparisons with the results for other NTE tests obtained in other studies, IOX computed majority indices in addition to average per-item scores. This highlights an important issue regarding changes in the methodology used. As people in the field continue to improve the procedures used to assess a test's job-relatedness, the results obtained for different tests using different procedures are not directly comparable. This is not a reason, however, to abandon efforts for improvement. As new methods are adopted, tentative judgments regarding the appropriateness of using the tests appraised with the new methods may have to be made until enough tests have been appraised with the new procedure to allow for comparisons among tests, unless, of course, the results for a particular test are so low that even temporary use of the test cannot be justified.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bowman, H. L. (1984). Technical report on validation of the National Teacher Examinations Core Battery and Specialty Area Tests and recommendations for performance standards for initial certification in Tennessee. Memphis: Memphis State University.

Coyle, K. K. (1988, April). An NTE-appraisal shift: From adequacy-of-preparation to job relevance. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Cross, L. H. (1982). Final report: Validation study of the National Teacher Examinations for certification of entry-level teachers in the state of Virginia, Phase I. Blacksburg: Virginia Polytechnic Institute and State University.

Cross, L. H. (1984, April). Validation of the National Teacher Examinations for certification of entry-level teachers in the Commonwealth of Virginia. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Cross, L. H. (1985). Validation of the NTE tests for certification decisions. Educational Measurement: Issues and Practice, 4(3), 7-10.

Cross, L. H. (1988, April). Reservations and recommendations regarding the validation of the NTE Program tests. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Echternacht, G. (1985). Content validity analysis of the NTE specialty area test in special education. Princeton, NJ: Educational Testing Service.

Echternacht, G. (1986). A study of the validity of the NTE specialty area tests for use in Pennsylvania. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1976a). Report on a study of the use of the National Teacher Examinations by the state of North Carolina. Princeton, NJ: Author.

Educational Testing Service. (1976b). Report on a study of the use of the National Teacher Examinations by the state of South Carolina. Princeton, NJ: Author.

Educational Testing Service. (1979). Report on a study of the use of National Teacher Examinations by the state of Mississippi. Princeton, NJ: Author.

Faggen, J. (1983). Report on a study of NTE Core Battery tests by the state of New York. Princeton, NJ: Educational Testing Service.

Faggen, J. (1985). Report on a study of selected NTE and GRE tests by the state of New York. Princeton, NJ: Educational Testing Service.

IOX Assessment Associates. (1983). Appraising the National Teacher Examinations for the state of Kentucky. Los Angeles: Author.

IOX Assessment Associates. (1987). A report of test-appraisal studies and standard-recommendation activities for four National Teacher Examinations for the state of South Carolina. Los Angeles: Author.

IOX Assessment Associates. (1988). A report of test-appraisal studies and standard-recommendation activities for three NTE tests for the state of South Carolina. Los Angeles: Author.

Kalisch, S. J. (1986). Report on a revalidation study of selected NTE specialty area tests for use by the state of South Carolina. Princeton, NJ: Educational Testing Service.

Madaus, G. F., & Pullin, D. (1987). Teacher certification tests: Do they really measure what we need to know? Phi Delta Kappan, 69(1), 31-38.

Masonis, E. (1984). Report on a study of selected NTE specialty area tests by the state of South Carolina. Princeton, NJ: Educational Testing Service.

McMillan, J. H. (1983, April). Content validation study of the National Teacher Examination Area Examinations by Virginia K-12 teachers and instructional supervisors. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Mehrens, W. A. (1987). Validity issues in teacher licensure tests. Journal of Personnel Evaluation in Education, 1, 195-229.

NTE Policy Council. (1985). Guidelines for proper use of NTE tests. Princeton, NJ: Educational Testing Service.

Poggio, J. P. (1985). Report on the validation studies of the National Teacher Examinations Core Battery Tests for certification of entry-level teachers in the state of Kansas. Lawrence, Kansas: Center for Educational Testing and Evaluation, University of Kansas.

Radbil, A. (1983). Report on a study of the NTE Core Battery tests by the state of New Mexico. Princeton, NJ: Educational Testing Service.

Rebell, M. A. (1986). Disparate impact of teacher competency testing on minorities: Don't blame the test-takers--or the tests. Yale Law and Policy Review, 4, 375-403.

Roth, R. (1982, June). Validation of the NTE: Arkansas style. Paper presented at the Educational Commission of the States Annual Conference on Large-Scale Assessment, Boulder, Colorado.

United States Department of Education, Office of Educational Research and Improvement. (1987). What's happening in teacher testing. Washington, D.C.: Author.

United States v. South Carolina, 445 F. Supp 1094 (D.S.C. 1977), aff'd 434 U.S. 1026 (1978).